HW2

1.

A) The feature set: {popularity, activeness, interaction}. Feature popularity can be defined as the number of friends a Facebook account has with range from 0 to the number of Facebook users. Feature activeness can be defined as the number of posts the user has posted within 1 month. Feature interaction can be defined as the number of comments the user received and also posted within 1 month. All of them have the range of non-negative integer. I pick those features because it can represent whether the user is active or/and has a lot of friends or/and interact a lot with his or her friends. Those features represents majority of the functionality of Facebook.

B) Metric: The smallest number of friends needed to

C) Height in feet: 0.8 to 8; weight in kilograms: 3 to 110; number of hairs in head: 0 to 100,000. No, it does not make sense to use Euclidean 3-space where we treat the three elements equally because the range of three elements are largely different and the extent of their correlation to the classification is different. For example, number of hairs is mostly likely to be much larger than the height and weight that makes number of hairs the dominant factor deciding the classification. I would design a weighted metric so that the factor that affects the classification most would have the most weight and for the factors that have relatively large numerical values, they should have less weight.

D) The edit distance metric can be used to determine the distance between two strands of DNA because it represents the min cost, the min change we need to make two DNA strands identical. The larger the edit distance, the less similar the two DNA strands are. However, this metric can be adapted in a way that reflects different weight of different edition. For example, substitution cost of bases should not be same weighted as the insertion and deletion cost, because for two strands of DNA, a small difference in their length might indicate a larger dichotomy than a large difference in individual bases.

3.

A) It takes around 13 hours to run mesure\_error on the data in wikipediatypo using the 3esl.txt dictionary. If we try the 64 different combinations, the time would be 64 \* 13 which is 823 hours. If we use 10-fold cross validation, the time would be 64\*13\*10 which is 8230 hours.

B)